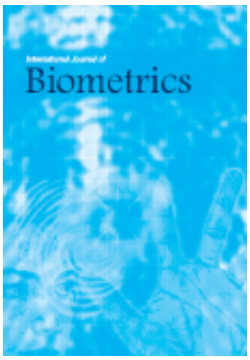


International Journal of Biometrics

ISSN online: 1755-831X - ISSN print: 1755-8301
https://www.inderscience.com/ijbm

Arabic offline writer identification on a new version of AHTID/MW database

Anis Mezghani, Monji Kherallah

DOI: 10.1504/IJBM.2024.10054549

Article History:

Received:	16 July 2022
Last revised:	28 November 2022
Accepted:	11 December 2022
Published online:	01 December 2023

Arabic offline writer identification on a new version of AHTID/MW database

Anis Mezghani*

Higher Institute of Industrial Management,
University of Sfax,
Sfax 3021, Tunisia
Email: anis.mezghani@gmail.com
*Corresponding author

Monji Kherallah

Faculty of Sciences,
University of Sfax,
Sfax 3000, Tunisia
Email: monji.kherallah@gmail.com

Abstract: Handwriting is considered to be one of the commonly used biometric modalities to verify and identify persons in commercial, governmental and forensic applications. In order to test and compare the accuracy of a computer vision system, in general, and a biometric system in particular, standard rich databases must be publicly available. In this paper and for this purpose, we expose the different works of writer identification of Arabic handwritten text carried out on our already published database AHTID/MW. As researchers have achieved high identification rates, we propose to extend the AHTID/MW database with new Arabic native writers and raise the level of difficulty. A baseline is drawn on each text-line image, and ground truth information is provided for each text image. In addition we present our experiments on the database using a new approach based on combining a CNN for feature extraction with GMM-based emission probability estimates for classification.

Keywords: Arabic writer identification; handwritten text image; AHTID/MW database; convolutional neural network; Gaussian mixture model; GMM.

Reference to this paper should be made as follows: Mezghani, A. and Kherallah, M. (2024) 'Arabic offline writer identification on a new version of AHTID/MW database', *Int. J. Biometrics*, Vol. 16, No. 1, pp.1–15.

Biographical notes: Anis Mezghani received his PhD in Computer Systems Engineering from the National Engineering School of Sfax, Tunisia. His research interests include image processing and analysis, pattern recognition and Arabic text recognition.

Monji Kherallah is a Professor at the Faculty of Sciences of Sfax, Tunisia. He obtained his PhD in Electrical Engineering in 2008 from the National Engineering School of Sfax, Tunisia. His research interests include applications of intelligent methods to pattern recognition and Arabic handwriting analysis and recognition. He is an IEEE member and a frequent reviewer for international journals.

1 Introduction

Biometrics is a rapidly growing branch of information technology. Biometric handwriting recognition and identification is built around the idea that every person incorporates individual features into his or her handwriting which can be used to distinguish the identity of the corresponding writer. Consequently, each writer can be characterised by his own handwriting by the reproduction of details and unconscious practices. This is why in certain cases of expertise handwriting samples can be used and considered instead of (or in addition to) fingerprints.

The individuality of writing style is the result of two main influential factors. The first is the taught method usually based in copybook that varies according to the geographical location, the temporal circumstances, and the cultural and historical backgrounds. The mechanical repeating of standard graphical templates such as letters, words and sentences, provides the basic individual writing style. The second is time; with the passage from childhood to adulthood, humans develop their own individual writing characteristics which are considered ever spontaneous. The net result is the handwriting process making possible to identify a person.

There are two major directions that take advantage of the analysis of handwriting. The first branch is graphology which is defined as the process of determining character traits from handwriting. The second branch deals with the forensic examination of handwriting where someone tries to determine who wrote a specific document. Due to its very large applicability, handwriting recognition has always dominated research in handwriting analysis. Writer recognition has received renewed interest in the last several years.

In the last score of years, most of the efforts in handwritten biometric recognition and identification have been focused on Latin script. This is due to the availability of several appropriate databases of handwritten texts (Nguyen et al., 2018; Obaidullah et al., 2018; Marti and Bunke, 2002; Iqbal et al., 2022; Grosicki et al., 2009; Brink et al., 2010; Singh et al., 2018). The existing research on writer identification and verification from handwritten Arabic text is still limited. The lack of freely available Arabic databases is considered as one of the main reasons of this limitation compared with other languages especially Latin. There are several applications for which writer identification and verification from Arabic handwriting text is important. Examples include forensic science and digital libraries including historical archives. Another interesting application consists to build personal handwriting recognition systems which are able to adapt themselves automatically to a particular writer in a multi-user environment. For that, researchers have prepared some databases for handwritten texts (Mahmoud et al., 2014; Maadeed et al., 2012) handwritten words (Pechwitz et al., 2002) and bank checks (Al-Ohali et al., 2003).

In this work, we describe the new version of the AHTID/MW database that covers all Arabic characters and forms (beginning, middle, end, and isolated). The proposed database contains Arabic words and text-lines written by 106 different writers. Two types of ground truths based on content information (text-line image and word image) are generated. The AHTID/MW database was then extended by a baseline ground truth annotation of text-line images. An earlier version of the database was presented in (Mezghani et al., 2013). Today, the number of writers is doubled and the database became larger with much more detailed description.

In Section 2, we will present the last version of AHTID/MW database and we will detail the used methods for preprocessing and data segmentation. In Section 3, we will outline the published works on writer identification related to the previous version of AHTID/MW. Section4 describes the proposed system based on a combination CNN-GMM approach. In the following Section, experimental results will be provided, opening thus the doors to future work on the developed database. Finally, concluding remarks will be given in Section 5.

2 AHTID/MW database description

This document complements previous reports on this database (Mezghani et al., 2013) which is acquired and tested by several international researchers/research groups and has been the subject of a competition in ICFHR conference (Mahmoud et al., 2014). The details and the initial statistics of the handwriting sample documents written by 53 individuals were reported in Mezghani et al. (2013). We report here the updated statistics on the database, the used preprocessing and segmentation methods of the paragraphs to lines and words, and details on the construction phase of the ground truth.

We describe now the extended version of the AHTID/MW database. The volunteers were first and foremost asked to provide information about the name, the age, the gender, and the level of education. The name is kept confidential in the ground truth file for reasons of confidentiality. Volunteers were then asked to handwrite four pages: the first and the second page contain 17 and 19 handwritten lines, respectively, to be copied by the 106 writers. Volunteers were then divided in two equal groups to complete the last two pages. The first group was asked to copy 17 lines for each page whereas the second group was asked to handwrite the same number of lines from the writer's imagination (or copied from a magazine or from whatever source). The first and the second pages are to be exploited for text-independent writer identification, whereas the third and the fourth pages are to be exploited for text-dependent writer identification. Figure 1 shows examples of such pages. To ensure some diversity, we have encouraged writers to change pen or pencil while writing. The handwritten texts have been scanned in greyscales using a 300 dpi resolution and stored in 'PNG' format.

2.1 Dataset Analysis and Statistics

The dataset contains a total of 424 scanned pages composed of approximately 44,730 words written by 106 writers of which approximately 432 words per writer for an independent analysis of the text and more than 10,386 words written by only 53 writers, who represent the half of the total number of writers, for a dependent analysis of the text. More details on the statistics of the AHTID/MW database are provided in Table 1.

Volunteer writers come from different education levels including high school and university students as well as employees, teachers, engineers and other jobs. Table 2 shows statistics on the age of writers. The number and the percentage of writers are provided for the whole of the database. The table shows that almost 70% of the texts were collected from writers aged 16 to 25. Indeed, most of the text was collected from university and high school students. Only three writers over 50 years old participated in this database. 4 of our writers are under the age of 15 and the majority of our pages were

written by writers aged between 15 and 25. The writers are also classified by their gender. Table 3 presents the number and the percentage of writers according to their sexes.

Figure 1 Format of documents per writer

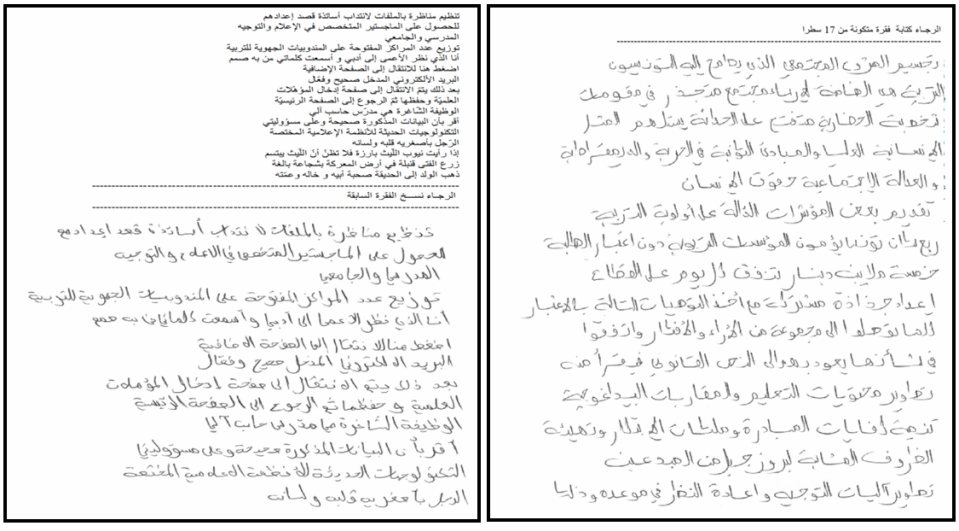


Table 1 Distribution of characters into different sets

	Sets for text-independent analysis				*Sets for text-dependent analysis*	
	Set 1	*Set 2*	*Set 3*	*Set 4*	*Set 3*	*Set 4*
Alif (ا)	8,586	10,494	5,300	5,883	5,018	5,635
Baa (ب)	2,226	2,544	1,113	1,007	1,106	1,113
Taaa (ت)	3,392	4,452	2,279	1,802	1,812	1,820
Thaa (ث)	424	424	106	159	139	172
Jiim (ج)	1,060	848	583	583	540	502
Haaa (ح)	1,696	1,272	318	477	636	506
Xaa (خ)	530	636	159	212	218	178
Daal (د)	1,908	1,696	901	848	867	726
Thaal (ذ)	636	318	212	212	250	233
Raa (ر)	2,544	2,650	1,537	1,749	1,163	1,620
Zaay (ز)	424	318	53	265	158	140
Siin (س)	1,272	1,272	689	795	533	661
Shiin (ش)	212	530	265	424	308	360
Saad (ص)	1,272	636	212	265	361	280
Daad (ض)	318	424	159	159	153	145
Thaaa (ط)	106	424	477	477	318	268
Taa (ظ)	742	530	106	159	322	258

Table 1 Distribution of characters into different sets (continued)

	Sets for text-independent analysis				*Sets for text-dependent analysis*	
	Set 1	*Set 2*	*Set 3*	*Set 4*	*Set 3*	*Set 4*
Ayn (ع)	2,226	2,438	1,060	954	1,030	896
Ghayn (غ)	424	212	53	424	70	253
Faa (ف)	1,272	1,166	424	636	520	414
Gaaf (ق)	742	2,120	583	742	467	600
Kaaf (ك)	742	848	530	742	318	512
Laam (ل)	8,374	7,420	3,498	3,922	3,302	3,633
Miim (م)	4,028	2,862	2,226	2,385	1,315	1,250
Nuun (ن)	2,332	2,332	1,431	1,378	1,378	1,333
Haa (ه)	1,590	1,908	583	636	953	798
Waaw (و)	2,014	2,756	1,749	1,007	1,401	1,101
Yaa (ي)	4,134	4,028	2,544	2,173	2,203	2,321
Hamza (ء)	0	106	159	0	130	178
HamzaAboveAlif (أ)	1,590	1,060	424	371	318	331
HamzaUnderAlif (إ)	742	636	424	424	149	256
TildAboveAlif (آ)	106	212	106	0	199	73
TaaaClosed (ة)	2,756	2,332	1,590	1,166	1,060	987
AlifBroken (ى)	424	318	0	53	76	85
HamzaAboveAlifBroken (ئ)	106	212	106	212	112	179
HamzaAboveWaaw (ؤ)	212	0	159	106	63	98
Quantity of characters	61,162	62,434	32,118	32,807	28,966	29,915
Quantity of PAWs	28,196	29,468	14,946	15,688	14,326	*14,624*
Quantity of words	11,448	11,448	5,459	5,989	5,314	*5,072*
Quantity of text-lines	1,802	2,014	901	901	901	901

Table 2 Statistics for age classification

Age	*<15*	*15–25*	*26–50*	*>50*
Number of writers	4	72	27	3
Percentage of writers	3.77	67.92	25.47	0.28

Table 3 Statistics for gender classification

Gender	*Number of writers*	*Percentage of writers*
Male	31	29.25
Female	75	70.75

2.2 Ground truth description

As the AHTID/MW database was intended to be used in Arabic sentence recognition, word recognition, word spotting, and writer identification, it seemed important to us to provide, with each folder containing handwritten samples, a ground truth data file which will ease substantially the accuracy assessment of any given approach or technique. The ground truth of the data was described in XML format. Figure 2 shows an example of an XML file describing the sequence of PAWs and characters making up the word 'المراكز'. A Latin character describing the character position in the PAW was added as a suffix for each character: 'I' stand for isolated, 'E' for end, 'M' for middle and 'B' for beginning character shapes. This is very important when using this database for word recognition or word spotting purposes.

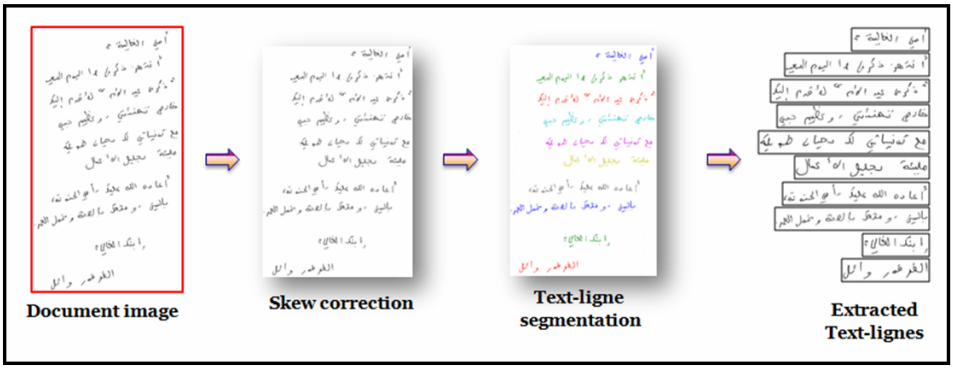
Figure 2 An example of a ground truth data files at the word level (see online version for colours)

```
- <wordImage id="15">
  - <content transcription="المراكز" nbPaws="4">
      <paw id="1" nbChars="1">Alif_I</paw>
      <paw id="2" nbChars="3">Laam_B Miim_M Raa_E</paw>
      <paw id="3" nbChars="1">Alif_I</paw>
      <paw id="4" nbChars="2">Kaaf_B Zaay_E</paw>
    </content>
  </wordImage>
```

2.3 Preprocessing and data segmentation

After the document images were scanned, text-line images were extracted automatically and stored in separate files. A noise reduction process and a skew correction (Kavallieratou et al., 2003) were applied before text-line segmentation. The text lines were identified using segmentation method proposed by Messaoud et al. (2012). We started by the application of the horizontal projection method. If the text lines are not well separated based on verification with the RLSA algorithm, we apply the grouping of connected components-based method for a second level of text line segmentation to detect the regions-of-problems. These represent the entry of a third segmentation method to correct overlap issues by the application of the nearest neighbour detection. As the spaces between the words define the word boundaries, we applied the vertical projections of the different components in the text line image. Those columns whose vertical projections exceeded a certain threshold were candidate words. The extracted lines and words were finally verified manually. An Overview of preprocessing and text-line multilevel segmentation process is shown in Figure 3.

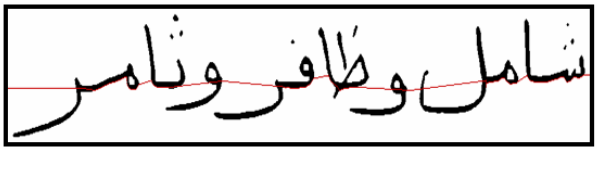
Figure 3 Overview of preprocessing and text-line segmentation process (see online version for colours)



2.4 Arabic handwritten baseline

A particular writer may be identified from writing characteristics such as form, slant, and baseline habits (Bouibed et al., 2021). The baseline is a vertical reference position for the characters and sub-words in a handwritten text-line image. The baseline can be used in skew normalisation to segment the Arabic text into words or characters (Fergani and Bennia, 2018) and to make the text ready for the feature extraction stage (Sueiras and Ru, 2018). Therefore, we have extended the AHTID/MW database by a baseline ground truth annotation of text-line images which enable researchers to take advantage of the writing line characteristic and to evaluate and compare their baseline detection algorithms with other published works. In Mezghani et al. (2013), we proposed a baseline extraction method to provide a fine baseline estimation of handwritten text-lines. The developed system contains three steps: the first is the detection and the removal of diacritics. The second step is to extract the upper and the lower baselines according to the horizontal projection histogram. In the last step, the baseline was estimated more precisely using support points. We evaluated the proposed method on AHTID/MW database and we obtained a percentage of 88.7% satisfaction using a threshold of seven pixels as an average pixel error. Figure 4 shows the baseline estimation of an Arabic handwritten text-line from AHTID/MW database.

Figure 4 Baseline estimation of an Arabic handwritten text-line (see online version for colours)



Source: Mezghani et al. (2013)

3 Related works on AHTID/MW

Research in writer identification and verification is still predominantly aimed for Latin and other western scripts; only limited number of researchers has addressed writer identification and verification of Arabic text (Mostafa et al., 2019). We present in this section the published writer identification works carried out on the last version of AHTID/MW published in Mezghani et al. (2012). Table 4 summarises writer identification state of the art approaches tested on AHTID/MW.

Table 4 Writer identification state of the art approaches on AHTID/MW

System	*Year*	*Feature source*	*Writer identification rate*
Slimane and Märgner (Slimane et al., 2014)	2014	Sliding window	69.48%
Schomaker and Bulacu implemented in Schomaker and Bulacu (2004)	2017	Contours	66.40%
Hannad et al. (2016)	2016	Fragments	77.30%
Khan et al. (2016)	2016	Full page	87.50%
Khan et al. (2017)	2017	Overlapping blocks	71.60%
Khan et al. (2019)	2019	Words	95.60%
Chahi et al. (2018)	2018	Connected components	99.53%
Chahi et al. (2019)	2019	Connected components	99.53%
Jaiem et al. (2014)	2014	text line	37.51%

A Gaussian mixture models-based system was developed by Slimane and Märgner (2014) for Arabic writer identification. A fixed-length sliding window technique was used on handwritten text line images to extract 21 features. The system provides log-likelihood scores. The Gaussian Mixture Model with the highest score is selected to identify the scripter. An accuracy of 69.48% has been reported using the AHTID/MW database.

Jaiem et al. (2014) proposed a writer Identification system which is participated in the Competition on Arabic Writer Identification held at ICFHR'2014 (Slimane et al., 2014). The system was applied on line level using steerable pyramids. The authors used the Back propagation Artificial Neural Network as a classifier. A normalised image with 1,024×1,024 size is created for each handwritten text line image, and steerable pyramid is applied with six orientations at seven levels. Based on the variance for each level, a feature vector is built. A three-layer feed-forward neural networks classifier is applied to evaluate the performance of the proposed system witch achieved an accuracy of 37.51% on AHTID/MW.

In Khan et al. (2016), the authors proposed a writer identification system based on multi-scale local ternary pattern histogram for multi-resolution analysis. The majority voting rule is used for classification: the writer having the majority votes at the various scales is most probably the author of the examined text. The developed system has been applied on the Arabic text line images of the AHTID/MW database and reported a writer identification rate of 87.5%.

Hannad et al. (2016) presented a writer identification system based on textural information extraction from small text fragments. Firstly, the handwritten text image is segmented using a 100×100 pixels window size into small fragments. Each one is fed subsequently to three feature extractors: local binary pattern, local ternary pattern and local phase quantisation. The last one gave the best results of writer identification on AHTID/MW and produced an accuracy of 77.3%.

In Khan et al. (2017), the authors implemented the universal codebook approach adopted by Schomaker and Bulacu (2004). They used a self organising map to cluster connected contour components features to get a universal codebook. For each image, a descriptor histogram is generated by using the obtained universal codebook. The use of this approach on the AHTID/MW leads to an identification rate of 66.40%.

A bagged discrete cosine transform-based system was proposed in Khan et al. (2017) for offline text independent writer identification. The main used techniques of the proposed system comprise discrete cosine transform for local descriptor computation, bagging and clustering for multiple vector quantisations, localised histograms of vector codes for structured writer representation, kernel discriminate analysis for dimensionality reduction and nearest centre rule for classification. An accuracy of 71.6% has been reported using the text-line images of the AHTID/MW database by using set 1, 2, 3 for training and set 4 for testing. The authors mention that, since the DCT features specify the images frequency content, the use of greyscales images from AHTID/MW is perfect. The greyscales images carry acceptable frequency information to decrease the inter-class similarity.

Chahi et al. (2018) proposed a learning-based method consisting on the use of the Hamming distance with 1-NN for classification. Block wise local binary count (BW-LBC) was used as feature descriptor applied on extracted connected components. The density of white pixels is computed within blocks of the writing sample. The authors compared the obtained results with those obtained using the handcrafted descriptors local binary patterns (LBP), local phase quantisation (LPQ) and local ternary patterns (LTP), commonly used in writer identification. The best identification rate (99.53%) over AHTID/MW database was achieved thanks to the BW-LBC descriptor.

Khan et al. (2019) presented an off-line writer identification system by using scale invariant feature transform (SIFT) and RootSIFT descriptors to represent text data. Gaussian mixture model (GMM) was used to construct, for every writer, a set of similarity and dissimilarity Gaussian mixture model based on SIFT and RootSIFT descriptors. Each GMM generates an intermediate prediction score by using weighted histogram technique, which after that fused by a linear function in order to obtain a final prediction score. The system was evaluated on AHTID/MW and reported an identification rate of 95.60%.

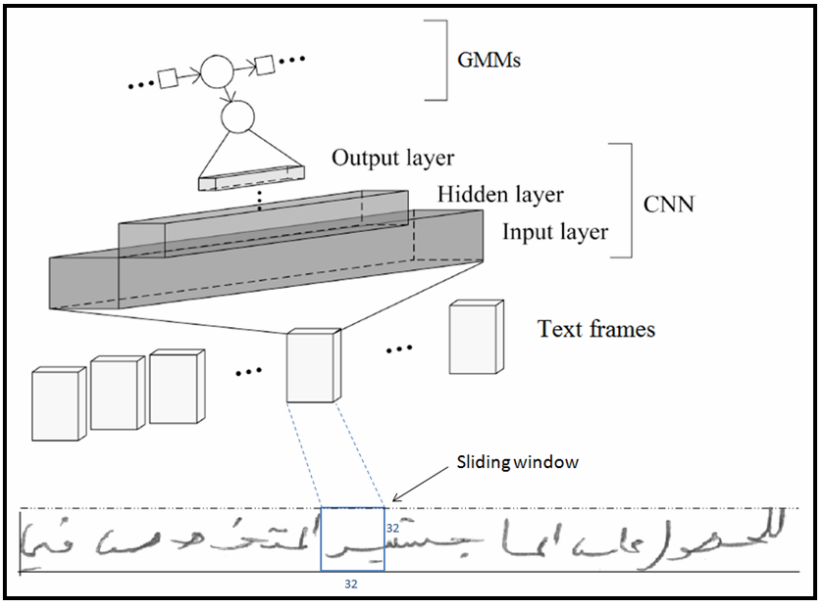
The same LBP, LPQ and LTP operators were used by Chahi et al. (2019) in the feature extraction phase. The authors have noticed that these operators perform worse when used directly in Chahi et al. (2018) as histogram descriptors. The new system is based on the technique of feature histogram construction and the dimensionality reduction. Each connected component is represented by a feature vector containing a set of computed histograms. The LPQ operator provides high identification by comparing with the tested descriptors.

4 GMM-CNN classification

In our previous papers (Noubigh et al., 2021; Mezghani et al., 2014), we used the AHTID/MW database for handwriting recognition and script identification. Hidden Markov models were used for the recognition purpose and GMM for script identification. These two techniques were also used in our paper (Mezghani et al., 2014) on complex documents from the MAURDOR database. In this paper, for writer identification purpose, we test firstly the implemented GMM-based system using the same features used in Mezghani et al. (2014). In a second time, we propose a novel approach based on combining a CNN for feature extraction with GMM-based emission probability estimates for classification. The proposed method exploits the representational capacity of CNN in the modelling of the variable appearance of writings. Meanwhile, GMM is used for the classification of the text sequence.

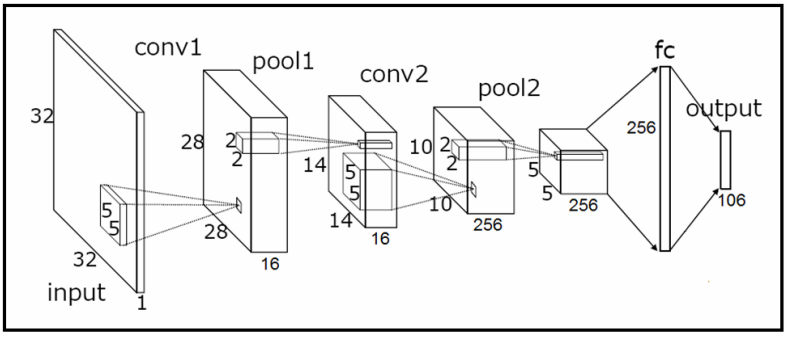
Deep CNN has shown good performance on various handwriting and identification tasks. CNN's efficiency results from the good ability to hierarchically learn features from a big amount of data. CNN is generally used as a static model whose input characteristic is fixed dimension. Generally, CNN is used in a static context. According to the literature, CNNs have not been widely used to identify writers. This is due to the disjunction of the test and training datasets for the same writer, making the CNN unable to classify the data. We therefore propose to apply CNNs to learn the activation features and GMMs for the classification step. To extract the sequence of text images of words or lines, we use the sliding window technique on the input image. The dynamic temporal model GMM is used to apply CNN to the sequence problem. Figure 5 presents the architecture of the proposed model.

Figure 5 Hybrid CNN-GMM architecture (see online version for colours)



In the proposed system, the CNN is used to extract features from text image windows. The back propagation is used for CNN training providing the ground truth label. The last layer is composed of 106 SoftMax nodes, which represent the different writers of the training set. After completion of training, the activation outputs are extracted from the first fully connected layer to represent the CNN features of the image frame. The CNN architecture of the proposed system is shown in Figure 6.

Figure 6 Architecture of the used CNN



The CNN architecture consists of two convolutional layers succeeded with pooling layers and two fully connected layers. The filter size of the first convolutional layer is 16 whereas that of second convolutional layer is 256. The max pooling is applied in the two pooling layers over regions of size 2×2. All these layers contain rectified linear units. The last layer is a 106-way softmax layer. These 106 nodes are used during the training process for classification. The training and test datasets consist of frame images of size 32 × 32 extracted by a fixed-length sliding window on the text-line image after normalisation with overlapping of one pixel. 256 CNN features represent each frame in the next step.

GMM are used for the computation of likelihood estimates of writer classes. The GMM is defined as the weighted sum of N component Gaussian densities as shown in the equation (1).

$$p(x|\lambda) = \sum_{i=1}^{N} \omega_i g\left(x|\mu_i, \sum_i\right) \tag{1}$$

with the following representations:

x D-dimensional feature vector

λ GMM model of the writer

ω_i N Gaussian mixture weights respecting the constraint $\sum_{i=1}^{N} \omega_i = 1$

$g(x \mid \mu_i, \sum_i)$ component Gaussian densities where $i = 1, \ldots, N$.

The GMM model is thus represented by $\lambda = \{\omega_i, \mu_i, \sum_i\}$ where $i = 1, \ldots, N$, μ_i represent the mean vector and $\sum_i$ represent the covariance matrix of the i^{th} component.

It is interesting now to estimate the GMM parameters defined by λ to find the best distribution of the training feature vectors. The maximum likelihood estimating technique is used to estimate these parameters. To maximise the GMM likelihood represented by equation (2), the expectation-maximisation (EM) algorithm (Dempster et al., 1977) is used.

$$p(X|\lambda) = \prod_{t=1}^{T} p\left(x_t|\lambda\right) \tag{2}$$

where $X = \{x_1, \ldots, x_T\}$ represent the training vectors and $t = 1, \ldots, T$.

5 Experimental results

We present in this section the experimental results carried out for Arabic writer identification using AHTID/MW database. We have tested the GMM-based identification system (Mezghani et al., 2017) and the GMM-CNN classifier described in Section 4. We used 5,618 lines for training and 1,802 lines for testing.

After conducting some experiments using the training set of the extended AHTID/MW database, we defined the optimal parameters of our network architecture. The filter size and the number of hidden nodes are fixed respectively to 5×5 and 256. Table 5 shows the obtained writer identification accuracy. Firstly, we tested the GMM-based system (A) developed for script identification (Mezghani et al., 2016, 2017). This system relies on handcrafted features. We obtained an accuracy of 57.05% with 256 Gaussian mixtures. In a second experiment (B), we used CNN for training and classification. We obtained an accuracy of 46.61% after 20 epochs of training.

We choose in our experiments 512 mixture components to train the hybrid CNN-GMM system (C). This value gives the best performance of writer category classification with an accuracy of 69.37% on test data. The identification accuracy of CNN-GMM increases by incrementing the number of mixture components and stops at 512. The continuous increase of the amount of Gaussian mixtures leads to the overfitting of the model.

Table 5 Writer identification results on AHTID/MW

System	*Used features*	*Classifier*	*Writer identification rate*
A	Handcrafted features (sliding window)	GMM	57.05%
B	Learned features (CNN)	CNN	46.61%
C	Learned features (sliding window + CNN)	GMM	69.37%

The hybrid model CNN-GMM improves the obtained performance of the GMM model by a margin of 12.32% and the performance of the CNN alone for training and classification by a margin of 22.76%. This demonstrates the interest of the combination developed in the present work and also the superiority of the deep model compared to the shallow model. The experiments show clearly that the hybrid CNN-GMM model increases significantly the performance of the GMM. It seems that CNN is replacing handcrafted features for a wide variety of issues.

By analysing the obtained results, it is clear that the writer identification results need a lot of improvement. Therefore, it seems that the automatic identification in Arabic

script is more difficult than several other scripts. Besides, the consideration of a large number of writers, which is necessary in the real world scenarios, complicates the task of Arabic writer identification. We can conclude that due to the specificities of the Arabic script, the identification of the Arabic writer is more difficult than the writer identification in other scripts. So, further research is important for more robust identification models.

6 Conclusions

In this paper, we have proposed an Arabic offline handwritten text images database written by 106 distinct writers from different age group, gender and qualification. A form composed of four pages was written by each writer and scanned at 300 dpi resolution. The database contains 7,420 text-line images and 44,730 word images. The images were partitioned into four sets to allow researchers to divide them into training and testing sets. A ground truth file is provided for each text image.

We also presented the work carried out by several international researchers/research groups on our first version of the database published in Mezghani et al. (2012) that has been the subject of a competition in ICFHR conference (Slimane et al., 2014). The database has already been used in writer identification and also in several other areas such as handwriting recognition and OOV detection, … Experiments on Arabic writer recognition were conducted using 7,420 segmented lines of AHTID/MW database and CNN-GMM classifiers. Our idea is mainly motivated by the complementary modelling power of CNN and GMM. Specifically, GMM is used for temporal modelling and CNN handles all sorts of writing variations. The CNN is important for its great capacity for representation and proves that it vastly outperforms several local features.

The comparatively low identification rates explain the importance of future research in this area and should encourage researchers to tackle this difficult task. We believe and hope that our database will be of great help and value for the research community. We invite all researchers who have worked also those who have not yet worked on the database to test their system with the entire database presented in this paper.

References

Al-Ohali, Y., Cheriet, M. and Suen, C. (2003) 'Databases for recognition of handwritten Arabic checks', *Pattern Recognition*, Vol. 36, pp.111–121, https://doi.org/10.1016/S0031-3203(02)00064-X.

Bouibed, M.L., Nemmour, H. and Chibani, Y. (2021) 'SVM-based writer retrieval system in handwritten document images', *Multimed. Tools Appl.*, https://doi.org/10.1007/s11042-020-10162-7.

Brink, A., Niels, R., van Batenburg, R., van den Heuvel, C. and Schomaker, L. (2010) 'Towards robust writer verification by correcting unnatural slant', *Pattern Recognition Letters*, Vol. 32, No. 3, pp.449–457, https://doi.org/10.1016/j.patrec.2010.10.010.

Chahi, A., Elkhadiri, I., Elmerabet, Y., Ruichek, Y. and Touahni, R. (2018) 'Block wise local binary count for off-line text-independent writer identification', *Expert Syst. Appl.*, Vol. 93, pp.1–14, http://dx.doi.org/10.1016/j.eswa.2017.10.010.

Chahi, A., Elmerabet, Y., Ruichek, Y. and Touahni, R. (2019) 'An effective and conceptually simple feature representation for off-line text-independent writer identification', *Expert Syst. Appl.*, Vol. 123, pp.357–376, https://doi.org/10.1016/j.eswa.2019.01.045.

Dempster, A., Laird, N. and Rubin, D. (1977) 'Maximum likelihood from incomplete data via the em algorithm', *Royal Statistical Society Series B Methodological*, Vol. 39, No. 1, pp.1–38, https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Fergani, K. and Bennia, A. (2018) 'New segmentation method for analytical recognition of Arabic handwriting using a neural-Markovian method', *International Journal of Engineering and Technologies*, Vol. 14, pp.14–30, https://doi.org/10.18052/www.scipress.com/IJET.14.14.

Grosicki, E., Carré, M., Brodin, J. and Geoffrois, E. (2009) 'RIMES evaluation campaign for handwritten mail processing', *Int. Conf. on Document Analysis and Recognition*, pp.941–945.

Hannad, Y., Siddiqi, I. and El Kettani, M.E.Y. (2016) 'Writer identification using texture descriptors of handwritten fragments', *Expert Systems with Applications*, Vol. 47, pp.14–22, https://doi.org/10.1016/j.eswa.2015.11.002.

Iqbal, M.A., Das, A., Sharif, O. et al. (2022) 'BEmoC: a corpus for identifying emotion in Bengali texts', *SN Computer Science*', Vol. 3, No. 135, https://doi.org/10.1007/s42979-022-01028-w.

Jaiem, F.K., Kanoun, S. and Eglin, V. (2014) 'Arabic font recognition based on a texture analysis', *International Conference on Frontiers in Handwriting Recognition*, pp.673–677, https://doi.org/10.1109/ICFHR.2014.118.

Kavallieratou, E., Dromazou, N., Fakotakis, N. and Kokkinakis, G. (2003) 'An integrated system for handwritten document image processing', *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 17, No. 4, pp.617–636, https://doi.org/10.1142/S0218001403002538.

Khan, F., Tahir, M.A., Khelifi, F. and Bouridane, A. (2016) 'Offline text independent writer identification using ensemble of multi-scale local ternary pattern histograms', *European Workshop on Visual Information Processing*, pp.1–6, https://doi.org/10.1109/EUVIP.2016.7764587.

Khan, F.A., Kheli, F., Tahir, M.A. and Bouridane, A. (2019) 'Dissimilarity Gaussian mixture models for efficient offline handwritten text-independent identification using sift and rootsift descriptors', *IEEE Transactions on Information Forensics and Security*, Vol. 14, No. 2, pp.289–303, https://doi.org/10.1109/TIFS.2018.2850011.

Khan, F.A., Tahir, M.A., Kheli, F., Bouridane, A. and Almotaeryi, R. (2017) 'Robust off-line text independent writer identification using bagged discrete cosine transform features', *Expert Systems with Applications*, Vol. 71, pp.404–415, https://doi.org/10.1016/j.eswa.2016.11.012.

Maadeed, S.A., Ayouby, W., Hassaïne, A. and Aljaam, J.M. (2012) 'QUWI: an Arabic and English handwriting dataset for offline writer identification', *International Conference on Frontiers in Handwriting Recognition*, pp.746–751, https://doi.org/10.1109/ICFHR.2012.256.

Mahmoud, S.A., Ahmad, I., Al-Khatib, W.G., Alshayeb, M., Parvez, M.T., Märgner, V. and Fink, G.A. (2014) 'KHATT: an open Arabic offline handwritten text database', *Pattern Recognit*, Vol. 47, pp.1096–1112, https://doi.org/10.1016/j.patcog.2013.08.009.

Marti, U. and Bunke, H. (2002) 'The IAM-Database: an English sentence database for offline handwriting recognition', *Int. J. on Document Analysis and Recognition*, Vol. 5, No. 1, pp.39–46, https://doi.org/10.1007/s100320200071.

Messaoud, I.B., Amiri, H., El Abed, H. and Märgner, V. (2012) 'A multilevel text line segmentation framework for handwritten historical documents', *Proceedings of the International Conference on Frontiers in Handwriting Recognition*, pp.513–518, https://doi.org/10.1109/ICFHR.2012.159.

Mezghani, A., Kanoun, S., Bouaziz, S., Khemakhem, M. and El-Abed, H. (2013) 'Baseline estimation in Arabic handwritten text-line: evaluation on AHTID/MW database', *International Conference on Pattern Recognition Applications and Methods*, pp. 430–434, https://doi.org/10.5220/0004218704300434.

Mezghani, A., Kanoun, S., Khemakhem, M. and El Abed, H. (2012) 'A database for Arabic handwritten text image recognition and writer identification', *International Conference on Frontiers in Handwriting Recognition*, pp.397–400, https://doi.org/10.1109/ICFHR.2012.155.

Mezghani, A., Slimane, F. and Kherallah, M. (2017) 'Writing type, script and language identification in heterogeneous documents', *Int. J. Intell. Syst. Technol. App,* Vol. 16, No. 3, pp.225–245, https://dx.doi.org/10.1504/IJISTA.2017.085358.

Mezghani, A., Slimane, F., Kanoun, S. and Kherallah, M. (2016) 'Window-based feature extraction framework for machine-printed/handwritten and Arabic/Latin text discrimination', *International Conference on Intelligent Computer Communication and Processing*, pp.329–335, https://doi.org/10.1109/ICCP.2016.7737168.

Mezghani, A., Slimane, F., Kanoun, S. and Märgner, V. (2014) 'Identification of Arabic/French handwritten/printed words using GMM-based system', *Proceedings of the Colloque International Francophone sur l'Écrit et le Document*, pp.371–374, https://doi.org/10.24348/sdnri.2014.CIFED-29.

Mostafa, M.A., Al-Qurishi, M. and Mathkour, H.I. (2019) 'Towards personality classification through Arabic handwriting analysis', in Visvizi, A. and Lytras, M. (Eds.) *Research & Innovation Forum 2019. RIIFORUM 2019. Springer Proceedings in Complexity*. https://doi.org/10.1007/978-3-030-30809-4_51.

Nguyen, H.T., Nguyen, C.T., Bao, P.T. and Nakagawa, M. (2018) 'A database of unconstrained Vietnamese online handwriting and recognition experiments by recurrent neural networks', *Pattern Recognition*, Vol. 78, pp.291–306, https://doi.org/10.1016/j.patcog.2018.01.013.

Noubigh, Z., Mezghani, A. and Kherallah, M. (2021) 'Densely connected layer to improve VGGnet-based CRNN for Arabic handwriting text line recognition', *International Journal of Hybrid Intelligent Systems*, pp.1–15, http://dx.doi.org/10.3233/HIS-210009.

Obaidullah, S.M., Halder, C., Santosh, K.C. et al. (2018) 'PHDIndic_11: page-level handwritten document image dataset of 11 official Indic scripts for script identification', *Multimed. Tools Appl.*, Vol. 77, pp.1643–1678, https://doi.org/10.1007/s11042-017-4373-y.

Pechwitz, M., Maddouri, S.S., Maergner, V., Ellouze, N. and Amiri, H. (2002) 'IFN/ENIT – database of handwritten Arabic words', *Proc. of Colloque International Francophone sur l'Écrit et le Document*, pp. 129–136.

Schomaker, L. and Bulacu, M. (2004) 'Automatic writer identification using connected-component contours and edgebased features of uppercase western script', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 6, pp.787–798, https://doi.org/10.1109/TPAMI.2004.18.

Singh, P.K., Sarkar, R., Das, N. et al. (2018) 'Benchmark databases of handwritten Bangla-Roman and Devanagari-Roman mixed-script document images', *Multimed. Tools Appl.*, Vol. 77, pp.8441–8473, https://doi.org/10.1007/s11042-017-4745-3.

Slimane, F. and Märgner, V. (2014) 'A new text-independent GMM writer identification system applied to Arabic handwriting', *International Conference on Frontiers in Handwriting Recognition*, pp.708–713, https://doi.org/10.1109/ICFHR.2014.124.

Slimane, F., Awaida, S., Mezghani, A., Parvez, M.T., Kanoun, S., Mahmoud, S.A. and Märgner, V. (2014) 'ICFHR2014 competition on Arabic writer identification using AHTID/MW and KHATT databases', *International Conference on Frontiers in Handwriting Recognition*, pp.797–802, https://doi.org/10.1109/ICFHR.2014.139.

Sueiras, J. and Ru, V. (2018) 'Offline continuous handwriting recognition using sequence to sequence neural networks', *Neurocomputing*, Vol. 289, pp.119–128, https://doi.org/10.1016/j.neucom.2018.02.008.